

## **ABSTRACT**

Cyber-attack, via cyberspace, targeting an enterprise's use of cyberspace for the purpose of disrupting, disabling, destroying, or maliciously controlling a computing environment/infrastructure; or destroying the integrity of the data or stealing controlled information. The state of the cyberspace portends uncertainty for the future Internet and its accelerated number of users. New paradigms add more concerns with big data collected through device sensors divulging large amounts of information, which can be used for targeted attacks. Though a plethora of extant approaches, models and algorithms have provided the basis for cyber-attack predictions, there is the need to consider new models and algorithms, which are based on data representations other than task-specific techniques. However, its non-linear information processing architecture can be adapted towards learning the different data representations of network traffic to classify type of network attack. In this paper, we model cyber-attack prediction as a classification problem, Networking sectors have to predict the type of Network attack from given dataset using machine learning techniques. The analysis of dataset by supervised machine learning technique (SMLT) to capture several information's like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments etc. A comparative study between machine learning algorithms had been carried out in order to determine which algorithm is the most accurate in predicting the type cyber-Attacks. We classify four types of attacks are DOS Attack, R2L Attack, U2R Attack, Probe attack. The results show that the effectiveness of the proposed machine learning algorithm technique can be compared with best accuracy with entropy calculation, precision, Recall, F1 Score, Sensitivity, Specificity and Entropy.

# TABLE OF CONTENTS

CHAPTER NO	TITLE	PAGE NO
	ABSTRACT	i
	LIST OF FIGURES	vi
	LIST OF TABLES	viii
<b>1</b>	<b>INTRODUCTION</b>	<b>1</b>
	1.1 Data Science	1
	<i>1.1.1 Data Scientist</i>	1
	1.2 Artificial Intelligence	2
	1.3 Natural Language Processing	3
	1.4 Machine Learning	3
	1.5 Objectives	4
	1.6 Project Goals	5
<b>2</b>	<b>LITERATURE SURVEY</b>	<b>6</b>
	2.1 Literature Survey	6
<b>3</b>	<b>AIM AND SCOPE OF THE INVESTIGATION</b>	<b>14</b>
	3.1 Project Proposal	14
	3.2 Scope of the Project	14
	3.3 Existing System	14
	3.3.1 Disadvantages	15
	3.4 Preparing Dataset	15
	3.5 System Study	15

3.5.1	Classification of Attacks	15
3.5.2	Attacks Summary	17
3.6	Feasibility Study	21
3.6.1	Data Wrangling	21
3.6.2	Data Collection	21
3.6.3	Preprocessing	21
3.6.4	Building the Classification model	21
3.6.5	Construction of a Predictive model	22
<b>4</b>	<b>EXPERIMENTAL OR MATERIALS AND METHODS, ALGORITHMS USED</b>	<b>23</b>
4.1	Project Requirements	23
4.1.1	Functional Requirements	23
4.1.2	Non-Functional Requirements	23
4.1.3	Environment Requirements	24
4.2	Software Description	24
4.2.1	Anaconda Navigator	24
4.2.2	Jupyter Notebook	26
4.2.3	Python	28
4.3	Python Libraries Needed	30
4.3.1	NUMPY Library	30
4.3.2	PANDAS Library	31
4.3.3	MATPLOTLIB Library	31
4.3.4	SCIKIT-LEARN Library	31
4.3.5	TKINTER	32
4.4	System Architecture	32
4.5	UML Diagrams	33
4.5.1	Class Diagram	33

4.5.2	Use Case Diagram	34
4.5.3	Workflow Diagram	35
4.5.4	Activity Diagram	36
4.5.5	Sequence Diagram	37
4.5.6	Entity Relationship Diagram	38
4.6	Algorithms Used	39
4.7	Comparing Algorithm with prediction in the form of best Accuracy result	44
4.8	Module Description	48
4.8.1	Data Validation process and Visualization	48
4.8.2	Prediction of DOS Attacks	54
4.8.3	Prediction of R2L Attacks	55
4.8.4	Prediction of U2R Attacks	56
4.8.5	Prediction of Probe Attacks	57
4.8.6	Prediction of Overall Network Attacks	58
4.8.7	GUI based prediction results of Network Attacks	59
<b>5</b>	<b>RESULTS AND DISCUSSION, PERFORMANCE ANALYSIS</b>	<b>61</b>
5.1	Performance Analysis	61
5.2	Discussion	64
<b>6</b>	<b>SUMMARY AND CONCLUSION</b>	<b>65</b>
6.1	Summary	65
6.2	Conclusion	65
6.3	Future Work	65
	<b>REFERENCES</b>	<b>66</b>

<b>APPENDIX</b>	67
A. SOURCE CODE	67
B. SCREENSHOTS	118

## LIST OF FIGURES

FIGURE NO	FIGURE NAME	PAGE NO
1.1	Process of Machine Learning	4
3.1	Process of data flow Diagram	22
4.1	Anaconda Navigator Interface	26
4.2	System Architecture	33
4.3	Class Diagram	33
4.4	Use Case Diagram	34
4.5	Workflow Diagram	35
4.6	Activity Diagram	36
4.7	Sequence Diagram	37
4.8	Entity Relationship Diagram	38
4.9	Logistic Regression	40
4.10	Decision Tree Classifier	41
4.11	Random Forest Classifier	42
4.12	Support Vector Classifier	43
4.13	Data Frame	49
4.14	Percentage level of protocol type	50
4.15	Comparison of service type and protocol type	51
4.16	Data Validation	53
4.17	Prediction of DOS Attacks	54
4.18	Prediction of R2L Attacks	55
4.19	Prediction of U2R Attacks	56
4.20	Prediction of Probe Attacks	57
4.21	GUI based prediction	60
5.1	Accuracy comparison of DOS Attack	61
5.2	Accuracy comparison of R2L Attack	61
5.3	Accuracy comparison of U2R Attack	62
5.4	Accuracy comparison of Probe Attack	62

5.5	Accuracy comparison of Overall Attack	63
6.1	GUI before input	118
6.2	GUI after giving input	118

## LIST OF TABLES

TABLE NO	TABLE NAME	PAGE NO
3.1	Attack Types Grouped to respective class	17
3.2	Description of Attacks	18



# CHAPTER-1 INTRODUCTION

## DOMAIN OVERVIEW

### 1.1 DATA SCIENCE

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.

The term "data science" has been traced back to 1974, when Peter Naur proposed it as an alternative name for computer science. In 1996, the International Federation of Classification Societies became the first conference to specifically feature data science as a topic. However, the definition was still in flux.

The term "data science" was first coined in 2008 by D.J. Patil, and Jeff Hammer Bacher, the pioneer leads of data and analytics efforts at LinkedIn and Facebook. In less than a decade, it has become one of the hottest and most trending professions in the market.

Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.

Data science can be defined as a blend of mathematics, business acumen, tools, algorithms and machine learning techniques, all of which help us in finding out the hidden insights or patterns from raw data which can be of major use in the formation of big business decisions.

#### **1.1.1 Data Scientist:**

Data scientists examine which questions need answering and where to find the related data. They have business acumen and analytical skills as well as the ability to mine, clean, and present data.

Businesses use data scientists to source, manage, and analyze large amounts

of unstructured data.

### **Required Skills for a Data Scientist:**

- **Programming:** Python, SQL, Scala, Java, R, MATLAB.
- **Machine Learning:** Natural Language Processing, Classification, Clustering.
- **Data Visualization:** Tableau, SAS, D3.js, Python, Java, R libraries.
- **Big data platforms:** MongoDB, Oracle, Microsoft Azure, Cloudera.

## **1.2 ARTIFICIAL INTELLIGENCE**

Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving.

Artificial intelligence (AI) is intelligence demonstrated by machines, as opposed to the natural intelligence displayed by humans or animals. Leading AI textbooks define the field as the study of "intelligent agents" any system that perceives its environment and takes actions that maximize its chance of achieving its goals. Some popular accounts use the term "artificial intelligence" to describe machines that mimic "cognitive" functions that humans associate with the human mind, such as "learning" and "problem solving", however this definition is rejected by major AI researchers.

Artificial intelligence is the simulation of human intelligence processes by machines, especially computer systems. Specific applications of AI include expert systems, natural language processing, speech recognition and machine vision.

AI programming focuses on three cognitive skills: learning, reasoning and self-correction.

**Learning processes.** This aspect of AI programming focuses on acquiring data and creating rules for how to turn the data into actionable information. The rules, which are called algorithms, provide computing devices with step-by-step

instructions for how to complete a specific task.

**Reasoning processes.** This aspect of AI programming focuses on choosing the right algorithm to reach a desired outcome.

**Self-correction processes.** This aspect of AI programming is designed to continually fine-tune algorithms and ensure they provide the most accurate results possible.

### **1.3 NATURAL LANGUAGE PROCESSING (NLP):**

Natural language processing (NLP) allows machines to read and understand human language. A sufficiently powerful natural language processing system would enable natural-language user interfaces and the acquisition of knowledge directly from human-written sources, such as newswire texts. Some straightforward applications of natural language processing include information retrieval, text mining, question answering and machine translation. Many current approaches use word co-occurrence frequencies to construct syntactic representations of text. "Keyword spotting" strategies for search are popular and scalable but dumb; a search query for "dog" might only match documents with the literal word "dog" and miss a document with the word "poodle". "Lexical affinity" strategies use the occurrence of words such as "accident" to assess the sentiment of a document. Modern statistical NLP approaches can combine all these strategies as well as others, and often achieve acceptable accuracy at the page or paragraph level. Beyond semantic NLP, the ultimate goal of "narrative" NLP is to embody a full understanding of commonsense reasoning. By 2019, transformer-based deep learning architectures could generate coherent text.

### **1.4 MACHINE LEARNING**

Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of